(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2006 (26.01.2006)

PCT

(10) International Publication Number
WO 2006/010022 A2

(51) International Patent Classification:
G06T 11/20 (2006.01)

(21) International Application Number:
PCT/US2005/024323

(22) International Filing Date: 8 July 2005 (08.07.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/585,923 8 July 2004 (08.07.2004) US

(71) Applicant *(for all designated States except US)*: VANDERBUILT UNIVERSITY [US/US]; 1207 17th Avenue South, Suite 105, Nashville, TN 37027 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: NI, Terri, T. [US/US]; 9237 Weston Drive, Brentwood, TN 37027 (US).

(74) Agent: CLARKE, Dennis, P.; Miles & Stockbridge RC, Suite 500, 1751 Pinnacle Drive, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW

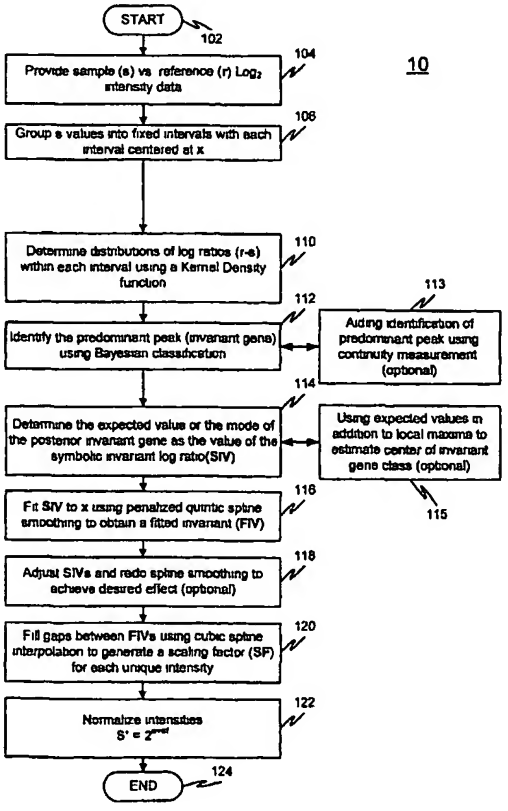
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: MICRO ARRAY DATA NORMALIZATION USING NONPARAMETRIC VARIABLE REDUCTION AND APPROXIMATION



(57) Abstract: A system and method for normalization of microarray data is provided. The method comprises a nonparametric variable reduction and approximation method with optional supervisibility of normalization and manual adjustability. The adjustability feature of the present invention gives a user full control of normalization and optimization. The method of and system of the present invention is generally suitable for use as a tool in any multivariate data analysis and, in particular, as a tool in microarray data analysis in particular.

WO 2006/010022 A2

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

MICROARRAY DATA NORMALIZATION USING NONPARAMETRIC VARIABLE REDUCTION AND APPROXIMATION

RELATED APPLICATIONS

[0001] This application claims the benefit under 35 U.S.C. § 119(e) to U.S. Provisional Application Serial Number 60/585,923, filed July 8, 2004, entitled "Microarray Data Normalization Using Nonparametric Variable Reduction And Approximation", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present invention relates generally to a method and system for analysis and identification of patterns in data collected from multivariate systems. In particular, the present invention relates to normalizing microarray data of genes, gene products, proteins, lipids, and combinations of the same so that patterns of interest in the data may be observed.

BACKGROUND OF THE INVENTION

[0003] Microarray is a powerful technology used to measure the expression of thousands of genes simultaneously. Experimental variability (noise) is inherent in microarray data and hampers the accurate estimation of true quantities that are measured. Using microarray technology for studying the developmental process in the whole embryo has provided insights on underlying molecular events. However, it poses additional challenges. The developmental process from one-cell to multi-organ formation involves alteration of expression in large amounts of genes. Although coexistence of biological variations and

experimental noises, and nonlinearity associated with both variables have been addressed, there are still some important issues to be resolved.

[0004] Current normalization methods can be divided into two categories: linear and intensity-dependent nonlinear normalization. Linear normalization relies on the assumption that noise structure is linear across the entire intensity range and, accordingly it may be appropriate to scale all data points by a single constant. Examples of this approach are median method, ratio statistics, Analysis of Variance (ANOVA) and centralization. One limitation of linear normalization techniques may be that they do not address nonlinear and intensity-dependent noise (intensity effect). To remove intensity effect in microarray data, various nonparametric approaches have been employed. Yang et al. applied the lowess plot smoother method to estimate the local center of log ratio over the local geometric mean space. Kelper et al. proposed an iteratively reweighted local regression algorithm to make inferences about slope and intercept that are intensity-dependent. Workman et al. fitted quantiles of two distributions using spline interpolation to achieve similar quantile distributions, while RMA proposed by Mzarray et al. directly equalized quantiles among multiple arrays. The above approaches require a critical biological assumption on gene effect: the number of genes that are up- and down- regulated is about the same. Although the rank invariant method to select genes that are not differentially expressed and outlier de-weighting strategy both improve robustness of normalization, the assumption is still required. To overcome this limitation, Fan et al. recently proposed a semilinear inslide model to assess gene effects from replications while removing intensity effects. However, the Fan method requires sufficiently large numbers of replicated arrays, and it does not address inter-slide intensity effect which may be important when aggregating gene effect information from replicated arrays.

[0005] Comparing microarray measurements of gene expression can reveal coregulation patterns that may be crucial in certain processes such as embryonic development. However, experimental variability (noise) in measured quantities can impede the quality of data comparison. Normalization may be required to remove systematic noise between any two arrays in comparison. A fundamental problem in normalization is distinguishing experimental variations from true biological variations.

SUMMARY OF THE INVENTION

[0006] The present inventors have discovered a normalization method, nonparametric valuable reduction and approximation (NVRA) using two-dimensional nonparametric modeling. Supervisibility is a unique feature of the method that gives a user full control of normalization and optimization. NVRA performs favorably as compared to other methods under the conditions tested, showing robustness in the presence of as much as 20% gene expression alteration, as well as an increase in the power of predicting differential gene expression, based on experimental data collected. NVRA may be particularly suitable to detect transcript level changes for developmental genes that are temporally and spatially expressed during embryonic development.

[0007] In an exemplary embodiment, the method of the present invention separates genes with biological variations from genes without biological variations and then adjusts experimental variations. An embodiment of the present invention removes the assumptions of gene effect and intensity effect that are often made in existing methods.

[0008] Further, in light of the above-mentioned limitations of conventional normalization methods, the present inventors conceived the NVRA method and system of the present invention. NVRA combines Bayesian kernel density classification and quintic spline approximation theory, and is applicable for data that contain nonlinear noise and large percentage of differentially expressed genes. The analysis method may be performed fully automatically or can be manually controlled and adjusted to address local difficulties or achieve desired results. This adjustability feature is not found in other normalization methods and is expected to give flexibility and control to a user of NVRA.

[0009] The normalization method presented here is precise, robust and well-controlled. Its effective reduction of systematic variability increases the sensitivity and specificity for detecting subtle biological changes between transcript mixtures of different cell types. NVRA is believed to be the first normalization algorithm that is nonparametric in two-dimensional space. In addition, the unique supervisibility feature enables a user to combine two different dynamic ranges in forming an extended one, or edit SIVs to optimize normalization such as correcting local error. These features allow NVRA to have broader application than existing methods. By contrast, the application of existing nonparametric methods is often conditional, requiring a symmetric distribution of differentially expressed genes, and conducted in unsupervised manners.

[0010] Experiments by the present inventors show that NVRA outperforms RMA and lowess under the condition tested here. Difference in performance may be due to NVRA's two-dimensional nonparametric feature, supervisibility and accuracy. It is known that kernel classifier performs better than linear discriminant analysis, and penalized quintic smoothing

spline provides smoother curve fitting (continuous up to the 4th degree derivatives), as compared to the first degree in lowess. The present inventors found that RMA and lowess are sensitive to local abnormal distribution present in some affymetrix spike-in chips, causing a visible spike in normalization curve, however, the performance of unsupervised NVRA did not appear to be affected. When the local predominant genes are not the invariant class as assumed, as the case of simulated data with 20% up-regulation (high intensity region), unsupervised NVRA does not normalize data correctly due to classification error. But the error can be corrected by supervised NVRA. Neither RMA nor lowess has such flexibility and ability to control normalization. Improved accuracy in normalization leads to more significant *P* values for true altered genes, but less significant *P* values for true unchanged genes. Both contribute to the increase in positive predictive values. A higher PPV, an indication of higher discovery rate, is expected to improve the concordance of results from independent microarray experiments.

[0011] Analysis of gene expression profiles of whole embryos during different developmental stages has been challenging. Many developmental regulatory genes control multiple cell types or lineages, and are expressed in a regional or asymmetric fashion in the whole embryo. Profiling of transcripts in whole embryos may fail to detect these subtle transcript level changes, especially if the expression level is low. The NVRA method provides signal-to-noise enhancement, and the success of our analysis using NVRA was immediately apparent. For example, most of genes that are known to be involved in zebrafish hematopoiesis and vasculogenesis were detected by our data analysis. Furthermore, our analysis of expression of these hematopoietic regulatory and functional components during

embryonic stages accurately reflected temporal expression pattern of the genes. The exemplary use of NVRA in analyzing zebrafish cloche embryos is discussed in greater detail below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG. 1 is a flowchart showing an exemplary embodiment of a nonparametric microarray data normalization method in accordance with the present invention;

[0013] FIG. 2 is a block diagram showing a system for nonparametric microarray data normalization in accordance with the present invention;

[0014] FIGS. 3a-3c are schematic views of exemplary NVRA normalization in a two-array setting;

[0015] FIGS 4a-4c show exemplary features of NVRA;

[0016] FIG. 5 shows exemplary performance of normalization methods;

[0017] FIGS. 6a-6e show exemplary *d o* phenotypes and defective gene expression;

[0018] FIGS 7a-7e show gene expression patterns in *d o* embryos versus wt siblings during development;

[0019] FIGS. 8a-8g show c/ø-dependent visual system defects;

[0020] FIG. 9 shows Positive Predictive Values (PPV);

[0021] FIG. 10 shows correlation of intensity between replicates;

[0022] **FIG.** 11 shows real-time PCR and profiling analysis;

[0023] FIGS. 12a- 12d shows graphs of exemplary effects of multimodality; and

[0024] FIG. 13 is a table showing a feature comparison of normalization methods.

DETAILED DESCRIPTION

OVERVIEW OF NVRA METHOD

[0025] The following is a general mathematical basis for the NVRA method and system.

[0026] Let x_t, y_t being the measured log2 sample and reference array intensity at the ith spot, or gene, or probe. Their relations are depicted in the following multivariate mixture model.

[0027] $$M_I = y_t - X_I = \delta_i + f(*,.) + \varepsilon_t, \qquad (1)$$

[0028] where δ_i denotes gene effect; $f(x_i)$ is a intensity dependent noise variable (intensity effect), and f_2 is heteroscedastic error variable whose variance is dependent on X_i.

[0029] To estimate intensity effect and gene effect independently, we first fixed the intensity effect by local modeling. Suppose at a local interval J centered at X_j, the data XJ e J is obtained if $|X_j - X_j| \leq h^*$, for $j = 1, 2, ..., m$. When h^* is small, X_j and the intensity effect, $f(x_j)$, are approximately constant for ally, then the equation (1) is rewritten as

[0030] $$M_j = \delta_j + f(x_j) + \varepsilon_j, \qquad (2)$$

[0031] where for all j, $f\{x_j\}$ is constant, and S_j is a random variable with mean of zero and a constant standard deviation. This step achieves variable reduction. For a 16 bit Affymetrix data and $h^* = 0.25$, about 37-43 of intervals are generated, duo to different dynamic ranges of sample intensity. The density distribution of M_j is estimated by the

nonlinear kernel density function. The modality of M_j distribution depends on the proportions of invariant genes ($\delta_j = 0$) and variant genes ($\delta_j \neq 0$). When the number of variant genes is negligible, M_j is unimodally distributed, and when the number is sufficiently large, variant genes form distribution pattem(s) distinguishable from invariant genes, causing multimodality.

[0032] Model (2) indicates that only when $\delta_j = 0$, the intensity effect $f(x_j)$ can be accurately assessed. We thus computed the posterior probability distribution of the invariant gene following Bayes Theorem, and its center as $f(x_j)$ at interval J. For a total of q numbers of intervals, the corresponding bivariate data set termed symbolic invariants $(x_q, f(x_q))$ was smoothed to remove arbitrary setting of interval width. The whole normalization procedure is outlined in FIG. 3.

Accuracy **of invariant gene selection**

[0033] We studies the accuracy of invariant gene selection using one simulated data where 20% of the genes are up- and down-regulated by 2 to 5 fold (up to down ratio is 1). The density distribution of M at one interval reveals multimodality feature. The SrVs from all intervals cover the entire intensity range (FIG. 4b, red) and align well with the expected line of normalization (FIG. 4b, green). By comparison, the invariants identified by the Rank Invariant method (FIG. 4c) were heterogeneous (black arrows) and biased (red arrow), and did not cover the entire intensity range. The results indicate that NVRA is more accurate than the Rank Invariant method for invariant gene selection.

Prevalence of asymmetric **distribution**

[0034] The major difference between NVRA and the existing nonlinear normalization methods is about the assumption of gene effect and density distribution pattern of $M_{.j}$ values. When there is no gene effect, the distribution is symmetric and unimodal, and the mean, median, and mode of the distribution are coincident. When the number of genes with gene effect increases, the distribution may become asymmetrical or multi-modal, leading to non-coincident of the mean, median and mode. Thus we could use the non-coincident as an indicator to quantify how often the assumption of a symmetric distribution is violated in real experimental datasets. The absolute deviation of mean or median of M_j from NVRA estimated $f(x_j)$ is computed. When the deviation was greater than 0.1, we arbitrarily define the distribution as mean or median non-coincident. The validity of the cutoff value 0.1 was confirmed after visual inspection of numerous distribution patterns.

[0035] For Affymetrix 59-chip dataset, a total of 58 normalizations were performed under default settings. 19 of them had at least one median non-coincident distribution. The total numbers of median and mean non-coincident distributions are 30, and 311 respectively. This corresponds to 1.36% and 14.09% of total 2208 distributions modeled respectively. For comparing wild type versus cloche at the same developmental stages, a total of 9 normalizations were performed under default settings. The numbers of median and mean non-coincident distribution are 0 and 11 respectively, out of 371 log ratio distributions. We did 2 normalizations across zebrafish developmental stages. The numbers of median and mean non-coincident distributions are 27 and 63 respectively. This corresponds to respective 31.4% and 73.26% of total 86 distributions modeled. The

results supported our prediction that when embryos develop from one cell to multi-cell organs, large numbers of genes are differentially expressed which causes the distribution of log ratios departing from symmetric unimodal distribution.

[0036] FIGS. 12a-12d illustrates examples of multimodality found in zebrafish data (FIG. 12a) and spike-in data set (FIG. 12b) and effect of multimodality on normalization (FIG. 12c,d). The absolute deviations of the median, mean from NVRA estimated values are 0.232, 1.726 (FIG. 12a) and 0.369, 0.229 (FIG. 12b). The results indicate that RMA and Lowess are sensitive to the presence of multimodal distributions while NVRA is not. FIGS. 12a-12b show a probability density distribution. The M values in FIG. 12a are from zebrafish cloche dataset #40 interval, sample = WT 24 hpf, reference = WT 15 hpf. The M values in FIG. 12b are from spike-in data set #27 interval, sample = chip 14, reference array = chip 25. The vertical lines in FIGS. 12a and 12b at the abscissa illustrate the center estimated by NVRA (green), mean (red) and median (black) approaches. FIGS. 12c-12d show a scatter plot of measured intensity (blue) superimposed by NVRA line (red), RMA line (green) and Lowess fitted values (cyan). s = chip 14, r = chip 25. Black arrows in FIGS 12c and 12d point to the cluster of data points that cause the multimodality shown in FIG. 12b.

[0037] Additional variables such as block or row effect can be added to the multivariate model proposed here. The same variable reduction and approximation strategy can be applied. It should be noted that the sequence of the variable reduction may give different results.

NVRA METHOD EXAMPLE

[0038] FIG. 1 is a flowchart showing an exemplary embodiment of a nonparametric microarray data normalization method 10 in accordance with the present invention. hi particular, the sequence begins at step 102 and continues to step 104.

[0039] IQ step 104, data is provided containing sample (s) versus reference (r) Log2 intensity. The sequence then continues to step 106.

[0040] In step 106, the s values are grouped into fixed width intervals. It should be appreciated that fixed-width intervals are used in this example for illustration purposes and that other intervals, including variable intervals, can be used. The sequence continues to step 108.

[0041] hi step 108, the mean of each interval (MI) is determined. The sequence continues to step 110. hi step 110, the distribution of the log ratios (r-s) within each interval are determined using a kernel density function. The sequence then continues to step 112.

[0042] hi step 112, the predominant peak (invariant gene) is identified using Bayesian classification. Optionally, in step 113, the identification of the predominant peak can be aided by using a continuity measurement. The sequence then continues to step 114.

[0043] In step 114, the center of the invariant gene log ratios is determined as the value of the symbolic invariant (SIV) log ratio corresponding to the maxima density. The mean density may also be used to identify the center. Optionally, in step 115, expected values may be used in addition to the local maxima to estimate the center of the invariant gene class. The sequence then continues to step 116.

[0044] In step 116, the SIV is fitted to the MI using penalized quintic spline smoothing to obtain a fitted invariant (FIV). The sequence continues to step 118.

[0045] hi step 118, which is optional, the automatic NVRA algorithm is able to be manually adjusted providing a supervisibility feature unique to NVRA. The SIVs can be adjusted and the spline smoothing can be reperformed to achieve the desired effect. The sequence continues to step 120.

[0046] hi step 120, the gaps between FIVs are filled using cubic spline interpolation to generate a scaling factor (SF) for each unique intensity.

[0047] In step 122, the intensities are normalized according to the formula $S' = 2^{s+sf}$. The sequence then continues to step 124, where the sequence ends.

NVRA SYSTEM EXAMPLE

[0048] FIG. 2 is a block diagram showing a system for nonparametric microarray data normalization in accordance with the present invention, hi particular, an NVRA system 20 comprises a data acquisition device 202, microarray data 204, normalization software 206, a processor 208, optional user input 210, normalized microarray data 212, a display 214, a printer 216, and optional other output devices 218.

[0049] hi operation, the data acquisition device acquires data, for example microarray data from a biological system. The microarray data 204 is transmitted, via a wired or wireless link (not shown) to the processor 208. The processor has access to the normalization software 206 that contains an implementation of an embodiment of the NVRA method of the present invention. Using the normalization software 206, the processor 208 performs normalization and optimization on the microarray data 204. The result is normalized microarray data 212, which may be transmitted, via a wired or

wireless link (not shown) to the display 214, the printer 216, or another output device 218. It should be appreciated that the system 20 shown is for illustration purposes only and that different components may be used and configured, distributed, connected and arranged according to a contemplated use of the invention. The connections may be over a network, such as the Internet, and may be wired or wireless. The processor may be software or hardware, or a combination.

[0050] While an embodiment of the NVRA method of the present invention is described below in relation to analysis of microarray data of zebrafish cloche mutant embryos. It should be appreciated that this example is provided for illustration purposes and that the NVRA method and system of the present invention can be used on any multivariate data and is not limited to microarray data or to the examples discussed below.

[0051] For example, to illustrate the usage of this technique, NVRA was applied to microarray data of zebrafish cloche (clo) mutant embryos that lack of blood and blood vessel formation. The analysis confirmed the reduced expression of known genes that are required at various developmental stages of hematopoiesis and vasculogenesis. Moreover, the analysis revealed a novel defective phenotype in cloche and detected reduced expression of a genetic network crucial for eye development. The follow-up biological experiments confirmed that the mutation in clo affects retinal development and causes blindness. Thus, the findings demonstrate the efficacy of NVRA, especially when applied in combination with microarray technology.

[0052] Genome-scale expression analysis has been widely used to investigate developmental regulation of gene expression in embryonic development and organ

formation. For example, the zebrafish cloche (clo) mutation causes defects in development of hematopoietic and endothelial lineages, resulting in loss of erythroids, myoloids and endothelial cells. The clo defect appears to be specific to the hematopoietic and vascular systems. Since positional cloning for identifying the clo gene has been hampered by the telomeric location of the clo locus, the mechanism and pathway whereby clo regulates hematopoietic stem cells and endothelial precursor cells remain enigmatic. In addition, it is not known whether clo affects development of other cell types and tissues.

[0053] The present inventors applied an embodiment of the NVRA method in analysis of transcriptional profiles of zebrafish clo mutant embryos at three key stages of embryonic development. The results indicate that the NVRA method is effective enough to recapitulate the global hematopoietic and endothelial defects in clo embryos. This analysis has led to the novel finding that expression of a series of eye-field transcription factors, as well as a family of G-protein-coupled photosensory opsin genes, is reduced in clo embryos. The visual system defects have been further validated by a series of analyses, including real-time RT-PCR, immunostaining, in situ hybridization and an optokinetic reflex assay. These data demonstrate that clo affects retinogenesis and visual function, in addition to previously reported hematopoietic and endothelial defects. Thus, our study represents an example that links development of a bio-informatic tool, namely the NVRA method of the present invention, to a novel biological finding.

NVRA METHOD EXAMPLE

[0054] FIGS. 3a-3dc show schematic views of exemplary NVRA normalization in a two-array setting. In particular, normalizing any sample array based on a reference

array according to the present invention requires that the raw intensities from the two arrays be logarithmically transformed. FIG. 3a is a schematic drawing showing generation of series of fixed-width intervals (r-s) along sample intensity values (s). FIG. 3b is a schematic plot revealing separation of log2 ratios (s-r) within an interval according to density distribution. FIG. 3c shows the generation of FIV and SF.

[0055] Coexistence of biological variation and experimental noise, as well as their respective nonlinearity, affects the fidelity of microarray data normalization. NVRA employs two major steps to resolve this complexity (see FIG. 3). The first step, variable reduction, identifies invariant genes (FIGS. 3 a, b). The entire sample intensity is divided into evenly spaced small intervals. The width is chosen to be small enough to have similar intensity values, but large enough to contain sufficient data points. Alternatively, a variable-width strategy could be employed to allow closer approximation in certain intensity regions. Within an interval, noise is approximately constant and the variation in log2 ratios (y-axis) is mostly due to differential expression between reference and sample array. NVRA delineates different forms of log2 ratio in a nonparametric manner. The resulting multi-modal distribution reflects different types of gene regulation. Typically three peaks are observed where the predominant peak corresponds to the invariant gene class, while the rest corresponds to the up- or down- regulated classes. In many cases, the predominant class is the peak with the highest density. However, when gene regulation events increase significantly, the existence of a predominant peak may not be obvious, then, the peak with the largest area under the curve is considered the invariant class. The symbolic invariant (SIV) is then derived by calculating the local maxima or the expected value of the invariant class.

[0056] The second step, variable approximation obtains, intensity-dependent scaling factor (SF) (FIG. 3c). The underlying nonlinearity in SIV represents relative noise between two arrays in each interval. The roughness in SIV is taken out by penalized quintic spline smoothing function in NVRA. The gaps between the interval-dependent fitted invariants (FIV) are then filled by cubic spline smoothing. The resulting SF represents the FIV as a function of sample intensity.

[0057] FIGS 4a-4c show exemplary features of NVRA as applied to identification of invariant genes. FIG. 4a shows multi-modal density (blue) distribution delineates various forms of log2 sample vs. reference ratios. FIG. 4b shows alignment of SIV (red) with theoretical SF values (green). FIG. 4c shows Invariant genes identified by the Rank invariant method (cyan). Blue highlights log2 intensity scatter plot of sample (synthetic data simulated with 20% up/down-regulation) vs. reference. (iid data) log (a-c). FIG. 4d shows supervised normalization to extend chip's dynamic range. SIV (red) and an unsupervised curve (green) are generated by automatic NVRA procedure. Black arrow points to the lowest SFV. The magenta curve represents supervised normalization values. The data from the supervised normalization was then used as the master reference array for NVRA normalization on any other chip of the spike-in dataset. Blue represents plotted x- and y- axis values.

[0058] A representative profile of variable reduction is shown (FIG. 4a) on simulated data where 20% of the genes are up- and down-regulated by 2 to 5 fold. The multi-modality reveals separation of the signal variation mixtures into distinct classes. Although all members of the invariant class can be taken for the subsequent approximation step, choosing a symbolic invariant (SIV) greatly reduces data

dimensionality. The SIVs from all intervals cover the entire intensity range (FIG. 4b) and all of them align well with the expected values (FIG. 4b). By comparison, the invariants identified by the Rank Invariant method (FIG. 4c) were heterogeneous, mixed with signal variants, and cover less than 50% of the intensity range. Adjusting parameters reduced the degree of heterogeneity, however, this change decreased the coverage as well. The results indicate that NVRA is more robust than the Rank Invariant method for variable reduction.

SUPERVISIBILITY OF NVRA

[0059] NVRA provides a way to control or redirect data normalization, which is not available in existing methods. The number of SIVs generated by NVRA is typically around 40 when the interval width is 0.25. This represents ~ 12,000 fold reduction in data dimensionality (or the number of data points) in the case of spike-in dataset. The manageable amount of SIVs allows human intervention or supervision so that the values of SIVs are modified to achieve a desired effect. One typical usage of this feature is to build a master reference array that has a longer dynamic range, as demonstrated in analyzing spike-in dataset (FIG. 4d). Unsupervised NVRA analysis generated a normalization curve (FIG. 4d) that curves up in two directions from the lowest point (FIG. 4d). From the lowest point to the right, normalization took place by adding increasingly larger values, leading to an increase in dynamic range; however, to the left, the scaling-up led to the shortening of dynamic range. To block this effect, we manually adjusted the SIVs on the left to be the value of SIV at the lowest point, and performed linear normalization on the corresponding region to generate a supervised normalization

curve (FIG. 4d). By preventing the shortening effect, the overall dynamic range of an array was extended.

NVRA PERFORMANCE

[0060] The present inventors performed two tests to evaluate and compare the performance of the NVRA, RMA and lowess methods. First, the fidelity of each normalization method was measured by using simulated data exhibiting marked variation in gene expression. The normalization error rate, defined as array mean of percent deviation of observed from true normalization factor, was used to directly measure the accuracy of normalization. Whereas a zero value indicates 100% accuracy in normalization, and a negative or positive value represents the under- or over-normalization error respectively. The percent of normalization error is plotted against percent of genes with altered gene expression.

[0061] FIGS. 5a-5d show an exemplary performance comparison of normalization methods. In FIGS. 5a-5c, the normalization error rate is plotted against the percentage of genes that are up-regulated (FIG. 5a), down-regulated (FIG. 5b), or up/down- regulated (FIG. 5c). FIG 5d shows a receiver operator characteristic (ROC) plot. Various true positive rates (y-axis) and the corresponding false positive rates (x-axis) were generated by gradually lowering the statistical *P* value thresholds. False positive rates greater than 0.05 are not shown here. Normalization by NVRA (red), RMA (blue), lowess (green) were applied on simulated data FIGS 5a-5c and affymetrix spike-in data FIG. 5d.

[0062] The difference between NVRA and other methods is very clear: NVRA is resistant to various degrees of gene regulation effect, while other methods are sensitive,

particularly to asymmetric gene regulation. The type of error incurred in RMA or lowess normalized data correlates well with the direction of simulated gene regulation. Up-regulation caused estimations to be shifted toward higher intensity values such that the normalized intensities are lower than the theoretical values (under-normalization, FIG. 5a). On the contrary, the presence of down-regulation resulted in over-normalization (FIG. 5b). hi the case of equal numbers of up- and down-regulated genes, data were under-normalized (FIG. 5c), but to a lesser degree than the up-regulation only, indicating that up-regulation has a larger effect than down-regulation.

[0063] The present inventors then tested and compared how each method improved the quality of inference in differential expression analysis of the Affymetrix human spike-in dataset, as used previously for evaluation and comparison of various normalization methods. Comparison of receiver operator characteristic (ROC) curves (FIG. 5d) indicated that NVRA achieved highest sensitivity (shown by the highest true positive at any given false positive), and highest specificity (1 - false positive). In the tests, the present inventors further calculated the positive predictive value (PPV) to measure the power of differential expression analysis (FIG. 9). FIG. 9 shows Positive Predictive Values (PPV). Data was generated as in FIG. 5d. Sensitivity: TP/(TP + FN). PPV: TP/(TP+FP). TP: true positive. FP: false positive. FN: false negative.

[0064] At 94% sensitivity, NVRA has PPV value of 60% while RMA and lowess are about 3.7%, indicating a 16-fold increase for NVRA. For achieving 98% sensitivity, NVRA is 22-fold more powerful than RMA and lowess. ROC scores reported here with RMA and lowess methods are comparable to the previously published data.

PROFILING OF ZEBRAFISH *CLO* EMBRYOS USING NVRA

[0065] As an example, the present inventors have applied the NVRA method to normalize gene expression data generated from zebrafish *d o* mutant embryos and sibling wild type (wt) embiyos at 15, 30 and 48 hour post fertilization (hpf). *d o* mutant embryos fail to generate hematopoietic stem cells in the region of intermediate cell mass (ICM), and fail to develop endothelial cells in the head and endocardium.

[0066] FIGS. 6a-6e show exemplary *d o* phenotypes and defective gene expression. FIG. 6a shows an anterior lateral view showing blood vessels expressing *EGFP* in the head of wild-type transgenic embryos *[Tg(β kl :EGFP)/+].* FIG. 6b shows an anterior lateral view revealing absence of blood vessels in the head of *d o* transgenic embryos *[do^{m39}/do^{m39}, Tgiflkl: EGFP)/+].* FIG. 6c shows a posterior lateral view showing hematopoietic stem cells expressing *EGFP* in the region of ICM in wild-type transgenic embryos *[Tg(gatal :EGFP)/+].* FIG. 6d shows a posterior lateral view revealing absence of hematopoietic stem cells in the ICM of *d o* transgenic embryos *[do^{m39}/do^{m39}, Tg(gatal:EGFPy+].* FIG. 6e shows the numbers of probe sets, including genes and ESTs, whose expression decrease or increase by more than 2-fold at embryonic stages 15, 30 and 48 hpf, respectively. In the figure, the blue arrow indicates endocardium, the yellow arrow indicates facial vessels, e indicates an eye, and the red arrow indicates hematopoietic stem cells.

[0067] Hematopoietic stem cells, marked by gatøi-driven *EGFP,* are completely absent in the ICM in *d o* transgenic embryos (FIGS. 6c-6d). *EGFP-* labeledy?A:7-positive endothelial cells are absent in the head and endocardium (FIGS. 6a-6b), but are detected in the trunk in *d o* transgenic embryos (data not shown). Thus, at 15 hpf, *d o* homozygous

embryos can be easily identified and collected for microarray analysis based on deficiency of *ga*toi-driven *EGFP* expression in the ICM (FIGS. 6c-6d). At 30 and 48 hpf, *do* homozygous embryos were recognized and collected based on absence of blood circulation and lack of blood vessels. Either experimental replicates (independent RNA samples) or technical duplicates (split before hybridization) were generated for each sample condition (FIG. 10).

[0068] FIG. 10 shows correlation of intensity between replicates. Technical duplicates: generated from the same RNA sample but by two independent hybridizations. # Experimental replicates: generated from independent RNA samples. +Pearson correlation coefficient (r) of intensity between replicates.

[0069] A total of 12 datasets of transcription profiles were generated using zebrafish Affymetrix GeneChips, containing 15,617 genes and expressed sequence tags (ESTs). The microarray data were normalized using NVRA. The normalized intensity has average Pearson correlation coefficient (r value) of 0.992 between replicates/duplicates (FIG. 10), indicating a high degree of reproducibility in sample collection, RNA preparation, and array hybridization. The present inventors selected genes that exhibit at least 2-fold expression difference with statistical significance (P value < 0.01) between *do* mutant and wt embryos. Overall, we identified 68, 109 and 1369 genes whose expression are decreased in *do* embryos at 15, 24 and 48 hpf, respectively. Likewise, 32, 524 and 373 transcripts whose expression levels are increased in *do* embryos were detected at each of three distinct stages, respectively (FIG. 6e).

[0070] Recapitulating global defects in hematopoietic and endothelial systems in *do* embryos

[0071] FIGS 7a-7e show gene expression patterns in *do* embryos versus wt siblings during development. FIG. 7a shows clustered expression patterns of 409 full

length genes at developmental stages 15, 30 and 48 hpf. Only genes that exhibited more than 2-fold changes ($P < 0.01$) are shown here. Each row represents one gene, and each column is a developmental stage. The scale represents log2 ratios. FIGS. 7b-7e show Log_2 expression ratios of *sd, gatal, Imo2, anddraciilin* (FIG. 7b); *hbbel, hbbe2, epb4.1, alas2, urod* andsptf/?(FIG. 7c); *mpx, lcpl, pbxla, pbx3b, p ιι 1* and *lyz* (FIG. 7d); *flkl,flil, tiel* and *ύe2* (FIG. 7e), in *d o* embryos versus wt embryos are plotted at various developmental stages.

[0072] The present inventors performed a hierarchical clustering analysis on 409 full-length mRNA whose expression are decreased or increased by more than two-fold at one of three developmental stages (FIG. 7a). A cluster of genes that displayed reduced expression in *d o* embryos was identified as genes important for hematopoietic and endothelial development. For example, transcription factors *sd, Imo2, gatal* and *draculin* play important roles in hematopoiesis, and it has been previously reported that expression of *sd, Imo2, gatal* is dependent on *do.* In the dataset of *d o* expression profiles, strongly reduced expression of *sd, Imo2, gatal* or *draculin* were observed at 15 and 48 hpf but to a lesser degree at 30 hpf in *d o* embryos. Furthermore, the reduced expression of these genes exhibited similar dynamic patterns (FIG. 7b). These data reflect that *d o* embryos are defective in two successive waves of primitive and definitive hematopoiesis at early and late stages. Thus, profiling of the whole embryo using NVRA algorithm is effective to detect temporal transcript level changes during embryonic development.

[0073] The present inventors next analyzed genes involved in development of erythroid, myeloid and endothelial lineages in a systematic manner in *d o* versus wt embryos. Many erythropoietic genes are required for heme and globin chain synthesis, including *hbbel,*

hbbe2, alas2 and *urod,* and for membrane stability, such as β-*spectiin (spβ)* and *epb4.1.* The dataset showed reduced expression of all these genes in *d o* embryos (FIG. 7c). Furthermore, the dataset revealed increased reduction of expression ratios of these genes in *d o* embryos from 30 to 48 hpf (FIG. 7c), consistent with increased expression of these genes during normal hematopoietic development. In wt embryos, transcription factor*pu. 1* controls myeloid differentiation and is expressed during myeloid development. In *d o* embryos, expression of the myeloid regulatory gene*pu.1* starts to decrease at 15 hpf onward, followed by significant reduction of myeloid-differentiated genes *mpx, lcpl,* and *lyz* at 30 hpf (FIG. 7d). At 48 hpf, expression of transcription factors*pbxla mdpbx3b* involved in megakaryocyte development, are reduced significantly. Endothelial *genesflkl, flil, tiel* and *tie2* are expressed at reduced levels in *d o* embryos at all developmental stages (FIG. 7e), consistent with biological observations. However, reduction *oïflil* expression is less severe at late stages compared to the early stage, suggesting that expression *offlil* in other non-endothelial tissues may not be decreased in *d o* embryos at late stages. Thus, the profiling analysis correlates well with developmental defects of hematopoietic and endothelial lineages in *d o* embryos. The dataset provides sufficient scope and accuracy to reflect global gene activity in *d o* and wt embryos.

IDENTIFYING CLO-DEPENDENT DEFECTS IN RETINAL DEVELOPMENT AND VISUAL FUNCTION

[0074] FIGS. 8a-8g show c/o-dependent visual system defects. In particular, FIG. 8a shows a pathway scheme representing an interactive network participating in eye development was generated by PathwayAssist software (Ariadne). All the genes shown here are orthologues or closest homologues of zebrafish genes that have reduced gene expression in *d o* at 48 hpf.

The purple line indicates physical interaction, the dotted line with arrowhead indicates transcriptional activation, the dotted line with vertical bar indicates transcriptional repression, and the blue line indicates coexpression. FIGS. 8b-8c show a dorsal view showing aggregation of melanosomes in the center of melanocytes in wt embryos (arrow in 8b), and a failure of aggregation of melanosomes in *d o* embryos (arrow in 8c). FIGS. 8d-8e show whole mount in situ hybridization analysis using *vsxl* as probe revealing normal expression of *vsxl* gene in wt retinae (arrow in FIG. 8d), and reduced *vsxl* expression in *d o* retinae (arrow in FIG. 8e) at 48 hpf. FIGS. 8f-8g show transverse sections with PKCα immunofluorescence showing normal patterning of bipolar cell in retinas of 5-day old wt larvae (arrow in FIG. 8f), and reduced number of bipolar cells in *d o* retinae siblings (arrow in FIG. 8g).

[0075] The comprehensive genomic view of transcription allows us to identify other genes affected by *do*. The dataset showed that expression of a group of eye-field transcription factors, including *pax ó, sine oculis homeobox3 (six3), visual system homeoboxl gene (vsxl), prospero-related homeobox gene (proxl), dadishundc {dachc)* and *meis2,* is reduced in *d o* embryos at 48 hpf (FIG. 8a). Together with *cone-rod homeobox gene (crx)* and photosensory *opsin* genes, these genes form a well-known interactive network that plays important roles in regulating retinal development and visual function (FIG. 8a). Indeed, expression of all five *opsin* genes on the GeneChip, including *opnlmwl, opnlmw2, opnlswl, opnlsw2* and *rhodopsin (rho),* as well as *crx,* is also reduced in *d o* embryos at 48 hpf (FIG. 8a). The decreased expression levels of *meis2, proxl, vsxl, dachc* and *opnlsw2* were verified by real-time RT-PCR (FIG. 11).

[0076] FIG. 11 shows real-time PCR and profiling analysis. Expression of *meis2, proxl, vsxl, dachc* and *opnlsw2* were examined using real-time PCR in *d o* and wt embryos at 30 and 48 hpf, and compared to the data from profiling analysis.

[0077] In situ hybridization analysis further validated the reduced *vsxl* mRNA levels in *d o* retinae as compared to wt siblings (FIGS. 8d-8e). Therefore, we examined whether retinal development and visual function are defective in *d o* embryos. Immunofluorescence analysis using *protein kinase C a* (PKCα) as a maker for bipolar cells revealed reduced number of bipolar cells in *d o* retinae compared to wt retinae (FIGS. 8f-8g). Furthermore, bipolar terminals in *d o* retinae only form one sublamina in the inner plexiform layer (FIG. 8g), while wt retinae develop two sublaminae of bipolar cell terminals (FIG. 8f). We next examined the visual function in *d o* embryos using the optokinetic assay. At 5 day post-fertilization, all tested wild-type (15/15) siblings displayed optokinetic response, while none of the *d o* embryos (0/15) responded to visual stimuli, suggesting that *d o* embryos are blind. The blindness phenotype of *d o* embryos is further supported by their failure to contract melanosomes under bright light, a vision-dependent reflex behavior (FIGS. 8b-8c). All together, these functional results demonstrate that *d o* affects retinal development and visual function, and provide biological validation for the results of microarray data analysis.

[0078] Most of genes that are known to be involved in zebrafish hematopoiesis and vasculogenesis were detected by our data analysis. Furthermore, our analysis of expression of these hematopoietic regulatory and functional components during embryonic stages accurately reflected temporal expression pattern of the genes. For example, in *d o* embryos, expression of hematopoietic regulatory genes *sd, gatal, lmo2,* and *draculin* is reduced significantly at 15 hpf,

whereas expression of eiythrocyte-functional genes *hbbel, alas2, urod, sptβ and epb4.1* is decreased dramatically at 48 hpf. Interestingly, the dataset recapitulated the second wave of down-regulation of expression of *scl, Imo2,* and *dracula* at 48 hpf in hematopoiesis. Endothelial genes, including^/, *tiel* and *tie2,* are normally expressed at lower levels than hematopoietic genes. In addition, in *d o* embryos, endothelial cells are dramatically reduced but not completely absent, whereas hematopoietic cells are completely absent. It therefore poses a challenge to detect alterations of endothelial transcripts when profiling the whole *d o* embryo. Reducing expression noise by NVRA analysis is sensitive enough for detecting a reduction *offlkl* expression in *d o* embryos with statistical significance.

[0079] In addition to confirming hematopoietic and endothelial defects, the results also suggested retinal defects in *d o* embryos. The dataset indicated that expression of a genetic network that important for eye development may be dependent upon the *d o* gene. This network includes a group of eye-field transcription factors, including *p ox ó, six3, vsxl,proxl, dachc* and *meis2* and a family of G-protein coupled photosensory *opsin* gene. The degrees of reduction of these genes predicted from our profiling analysis correlated well with those derived from real-time PCR tests. These data have led us to identify retinal defects and blindness phenotype in *d o* eyes using a series of biological experiments. Since *d o* mutation has been previously reported to affect hematopoietic and endothelial systems, this finding may bring us a new understanding of the role of *do.* These data suggest that *d o* is a critical regulatory gene that may control development of multiple lineages, including retinal, hematopoietic and endothelial cells. The retinal defects are likely not secondary to the absence of blood flow, because zebrafish mutant *moonshine* with absence of blood formation has normal retinal development and visual function. Since endothelial differentiation controls

development of certain adjacent organs, it is possible that the retinal defect in *do* embryos may be due to lack of endothelium. Alternatively, *do* may directly regulate retinal differentiation and development. In zebrafish, retinal differentiation occurs in three 10-hour-long consecutive waves beginning around 30 hpf, and most cells in the inner nuclear layer of zebrafish retina have just exited cell cycle and begun differentiation at 48 hpf. The reduced expression of the retinal genes in *do* mutant embryos occur at the onset of retinal differentiation, suggesting a potentially direct role of *do* in retinogenesis. In mouse, *vsxl* regulates retinal bipolar cell differentiation and visual pathway. The decreased *vsxl* expression in *do* retina at 48 hpf are likely responsible for the reduced bipolar cells. Thus, identification of *do* as a regulator of retinal development may provide another clue to how it may control the development of multiple lineages.

[0080] The method and system for nonparametric variable reduction and approximation, as shown in the above figures, may be implemented on a general-purpose computer, a special-purpose computer, a programmed microprocessor or microcontroller and peripheral integrated circuit element, an ASIC or other integrated circuit, a digital signal processor, a hardwired electronic or logic circuit such as a discrete element circuit, a programmed logic device such as a PLD, PLA, FPGA, PAL, or the like. In general, any process capable of implementing the functions described herein can be used to implement method and system for nonparametric variable reduction and approximation according to this invention.

[0081] Furthermore, the disclosed method and system for nonparametric variable reduction and approximation may be readily implemented in software using object or object-oriented software development environments that provide portable source code

that can be used on a variety of computer platforms. Alternatively, the disclosed method and system for nonparametric variable reduction and approximation may be implemented partially or fully in hardware using standard logic circuits or a VLSI design. Other hardware or software can be used to implement the systems in accordance with this invention depending on the speed and/or efficiency requirements of the systems, the particular function, and/or a particular software or hardware system, microprocessor, or microcomputer system being utilized. The method and system for nonparametric variable reduction and approximation illustrated herein can readily be implemented in hardware and/or software using any known or later developed systems or structures, devices and/or software by those of ordinary skill in the applicable art from the functional description provided herein and with knowledge of the computer and electronic test arts.

[0082] Moreover, the disclosed method and system for nonparametric variable reduction and approximation may be readily implemented in software executed on programmed general-purpose computer, a special purpose computer, a microprocessor, or the like. In these instances, the method and system for nonparametric variable reduction and approximation of this invention can be implemented as a program embedded on a personal computer such as a JAVA® or CGI script, or with another tool, such as Matlab, as a resource residing on a server or graphics workstation, as a routine embedded in a dedicated encoding/decoding system, or the like. The system can also be implemented by physically incorporating the method and system for nonparametric variable reduction and approximation into a software and/or hardware system, such as the hardware and software systems of bioinformatics systems, microarrays, and/or other biological lab equipment.

[0083] It is, therefore, apparent that there is provided in accordance with the present invention, a method and system for nonparametric variable reduction and approximation. While this invention has been described in conjunction with a number of embodiments, it is evident that many alternatives, modifications and variations would be or are apparent to those of ordinary skill in the applicable arts. Accordingly, applicants intend to embrace all such alternatives, modifications, equivalents and variations that are within the spirit and scope of this invention.

CLAIMS

1. A method for microarray data normalization using nonparametric variable reduction and approximation, said method comprising:

providing sample data having one or more intensity values;

grouping sample data into one or more intervals;

deteraiining a distribution of log ratios within each interval;

identifying a predominant peak in the distribution of log ratios corresponding to an invariant gene class;

determining the center of the invariant gene class as the value of the symbolic invariant log ratio corresponding to a local maxima of the invariant gene class;

fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant;

filling any gaps between fitted invariants using cubic spline interpolation to generate a scaling factor for each intensity value; and

producing normalized data by normalizing each intensity value using the scaling factor.

2. The method of claim 1, wherein the step of determining the distribution of log ratios within each interval includes applying a kernel density function;

3. The method of claim 1, wherein the step of identifying a predominant peak in the distribution of log ratios corresponding to an invariant gene class includes Bayesian classification.

4. The method of claim 1, wherein the step of identifying a predominant peak in the distribution of log ratios includes aiding identification of the predominant peak with continuity data.

5. The method of claim 1, wherein the step of determining the center of the invariant gene class includes estimating the center of the invariant gene class with expected values in addition to local maxima.

6. The method of claim 1, further comprising adjusting the symbolic invariant and re-fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant that achieves a desired effect.

7. The method of claim 6, wherein the step of adjusting the symbolic invariant is performed manually.

8. A computer program product for enabling a computer to normalize multivariate data, said computer program product comprising:

software instructions for enabling the computer to perform predetermined operations; and

a computer readable medium bearing the software instructions;

the predetermined operations including the steps of:

receiving sample data having one or more intensity values;

grouping sample data into one or more intervals;

determining a mean for each interval;

determining a distribution of log ratios within each interval;

identifying a predominant peak in the distribution of log ratios corresponding to an invariant gene class;

determining the center of the invariant gene class as the value of the symbolic invariant log ratio corresponding to a local maxima of the invariant gene class;

fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant;

filling any gaps between fitted invariants using cubic spline interpolation to generate a scaling factor for each intensity value; and

producing normalized data by normalizing each intensity value using the scaling factor, whereby the computer normalizes multivariate data.

9. The computer program product of claim 8, wherein the step of determining the distribution of log ratios within each interval includes applying a kernel density function;

10. The computer program product of claim 8, wherein the step of identifying a predominant peak in the distribution of log ratios corresponding to an invariant gene class includes Bayesian classification.

11. The computer program product of claim 8, wherein the step of identifying a predominant peak in the distribution of log ratios includes aiding identification of the predominant peak with continuity data.

12. The computer program product of claim 8, wherein the step of determining the center of the invariant gene class includes estimating the center of the invariant gene class with expected values in addition to local maxima.

13. The computer program product of claim 8, further comprising adjusting the symbolic invariant and re-fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant that achieves a desired effect.

14. The computer program product of claim 13, wherein the step of adjusting the symbolic invariant is performed manually in response to input from a user.

15 The computer program product of claim 8, wherein the multivariate data is microarray data.

16. A computer system for normalizing microarray data, said computer system comprising:

a processor; and

a memory including software instructions adapted to cause the computer system to perform the steps of:

receiving sample data having one or more intensity values;

grouping sample data into one or more fixed-width intervals;

determining a mean for each interval;

determining a distribution of log ratios within each interval by applying a kernel density function using Bayesian classification;

identifying a predominant peak in the distribution of log ratios corresponding to an invariant gene class;

determining the center of the invariant gene class as the value of the symbolic invariant log ratio corresponding to a local maxima of the invariant gene class;

fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant;

filling any gaps between fitted invariants using cubic spline interpolation to generate a scaling factor for each intensity value; and

producing normalized data by normalizing each intensity value using the scaling factor

17. The computer system of claim 16, wherein the step of identifying a predominant peak in the distribution of log ratios includes aiding identification of the predominant peak with continuity data.

18. The computer system of claim 16, wherein the step of determining the center of the invariant gene class includes estimating the center of the invariant gene class with expected values in addition to local maxima.

19. The computer system of claim 16, further comprising adjusting the symbolic invariant and re-fitting the symbolic invariant to the interval mean using penalized quintic spline smoothing to obtain a fitted invariant that achieves a desired effect.

20. The computer system of claim 19, wherein the step of adjusting the symbolic invariant is performed manually in response to input from a user.

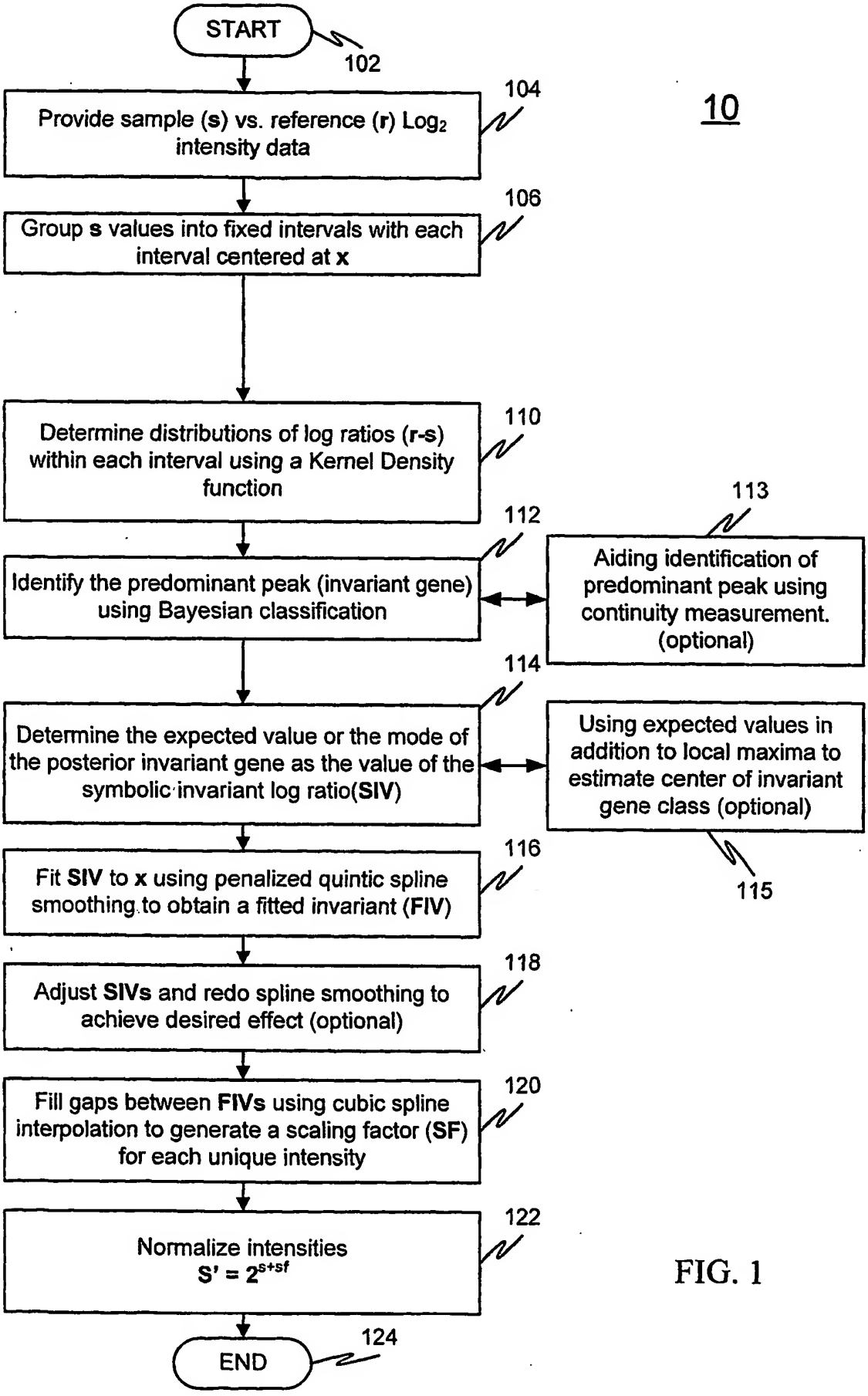

START
102
Provide sample (s) vs. reference (r) Log2 intensity data
104
10
Group s values into fixed intervals with each interval centered at x
106
Determine distributions of log ratios (r-s) within each interval using a Kernel Density function
110
113
Aiding identification of predominant peak using continuity measurement. (optional)
112
Identify the predominant peak (invariant gene) using Bayesian classification
114
Determine the expected value or the mode of the posterior invariant gene as the value of the symbolic invariant log ratio(SIV)
Using expected values in addition to local maxima to estimate center of invariant gene class (optional)
115
116
Fit SIV to x using penalized quintic spline smoothing to obtain a fitted invariant (FIV)
118
Adjust SIVs and redo spline smoothing to achieve desired effect (optional)
120
Fill gaps between FIVs using cubic spline interpolation to generate a scaling factor (SF) for each unique intensity
122
Normalize intensities
S' = 2s+sf
124
END

FIG. 1

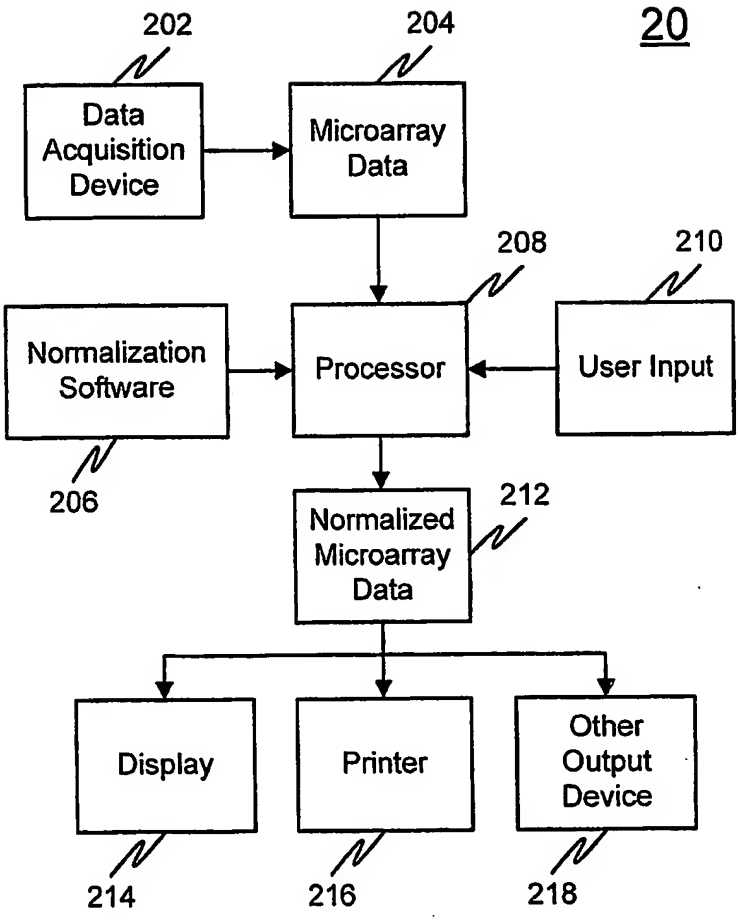

202
204
20
Data Acquisition Device
Microarray Data
208
210
Normalization Software
Processor
User Input
206
212
Normalized Microarray Data
Display
Printer
Other Output Device
214
216
218

FIG. 2

FIG. 3

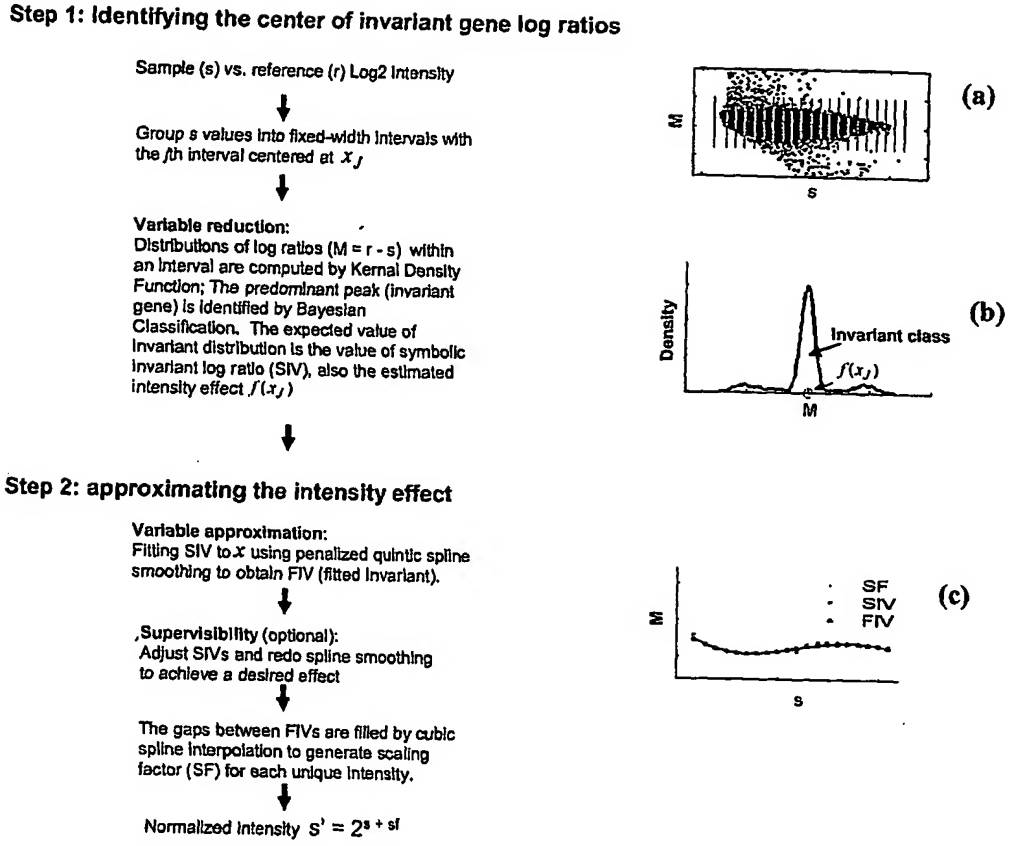

Step 1: Identifying the center of invariant gene log ratios
Sample (s) vs. reference (r) Log2 Intensity
Group s values into fixed-width intervals with the jth interval centered at x_j
Variable reduction:
Distributions of log ratios (M = r - s) within an interval are computed by Kernal Density Function; The predominant peak (invariant gene) is identified by Bayesian Classification. The expected value of invariant distribution is the value of symbolic invariant log ratio (SIV), also the estimated intensity effect $f(x_j)$
Step 2: approximating the intensity effect
Variable approximation:
Fitting SIV to x using penalized quintic spline smoothing to obtain FIV (fitted invariant).
Supervisibility (optional):
Adjust SIVs and redo spline smoothing to achieve a desired effect
The gaps between FIVs are filled by cubic spline interpolation to generate scaling factor (SF) for each unique intensity.
Normalized intensity $s' = 2^{s+sf}$
M
s
(a)
Density
Invariant class
$f(x_j)$
M
(b)
SF
SIV
FIV
M
s
(c)

FIG. 4

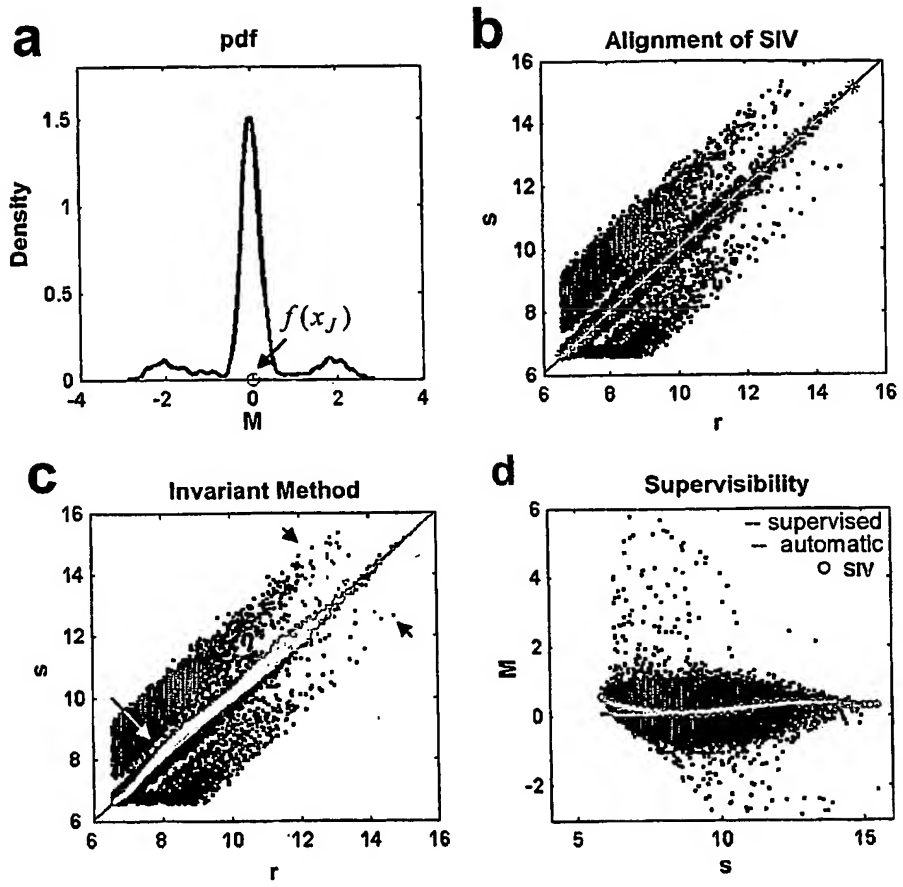

a
pdf
Density
M
$f(x_J)$
b
Alignment of SIV
s
r
c
Invariant Method
s
r
d
Supervisibility
supervised
automatic
SIV
M
s

FIG. 5

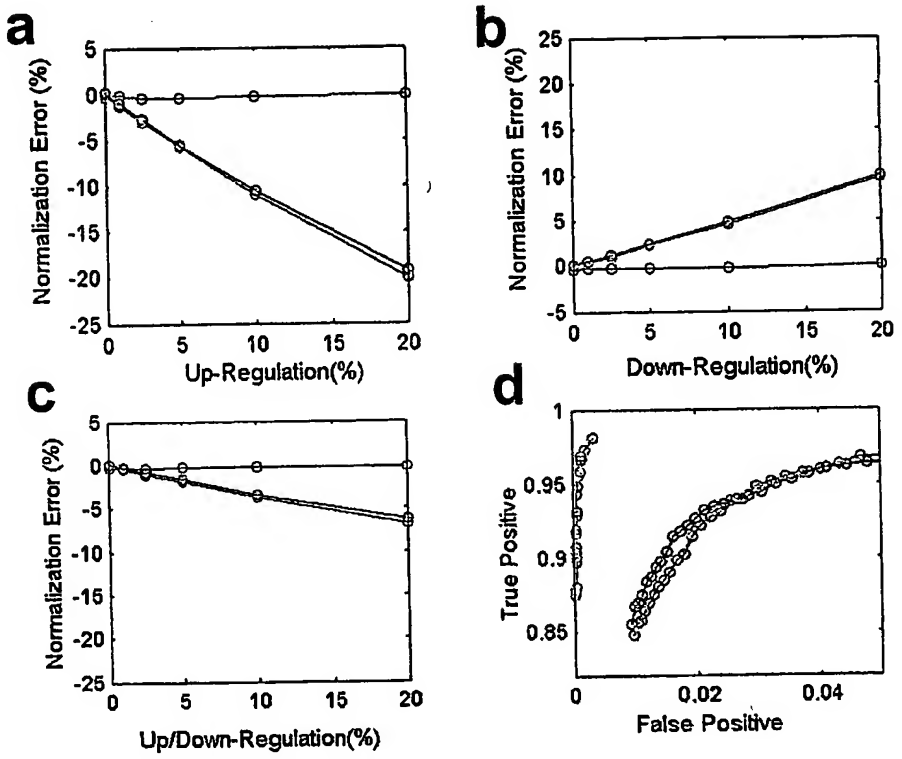


FIG. 9

Positive Predictive Value (PPV)

Sensitivity	PPV			Fold Increase
	NVRA	RMA	lowess	(NVRA/RMA)
0.9	0.7	0.055	0.064	13
0.92	0.66	0.048	0.054	14
0.94	0.6	0.037	0.038	16
0.96	0.45	0.026	0.026	17
0.98	0.26	0.012	0.01	22

FIG. 6

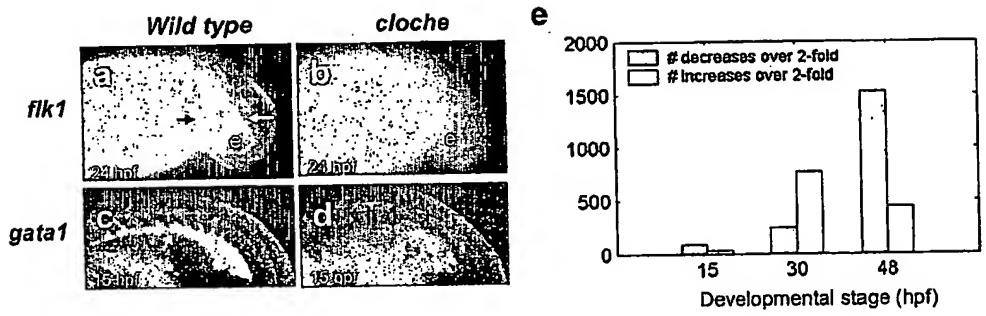


FIG. 10

Correlations between Replicates of Transcription Profiles

Stage	Number of Replicates	r Values+
15 hpf-*wt*	2	0.9906
15 hpf-*clo*	2	0.9874
24 hpf-*wt*	2	0.9957
24 hpf-*clo*	2	0.9955
48 hpf-*wt*	2	0.9918
48 hpf-*clo*	2	0.9897

+ correlation coefficeint

FIG. 7

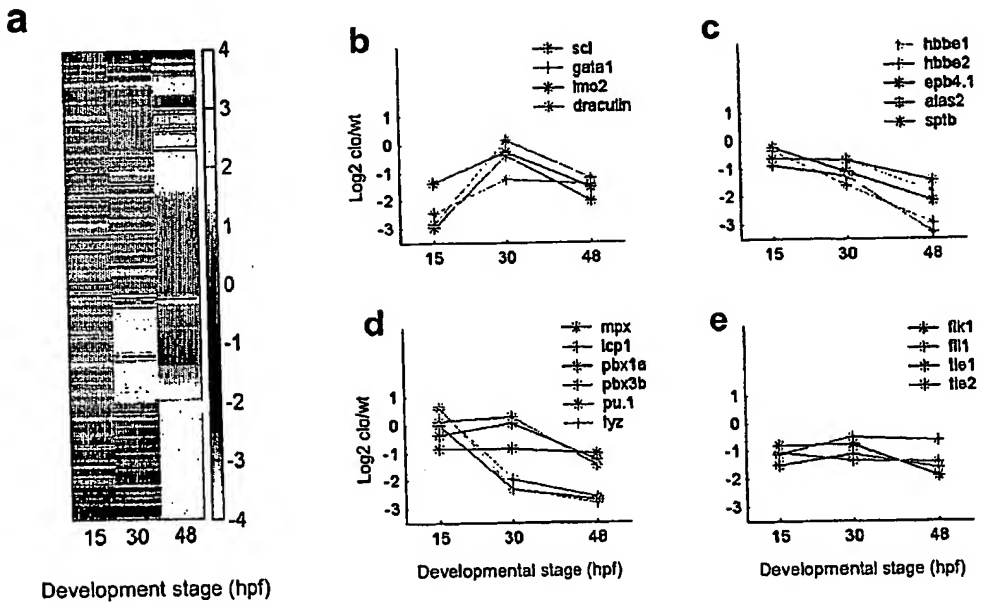

a
4
3
2
1
0
-1
-2
-3
-4
15 30 48
Development stage (hpf)
b
scl
gata1
lmo2
draculin
Log2 clo/wt
1
0
-1
-2
-3
15
30
48
c
hbbe1
hbbe2
epb4.1
alas2
sptb
d
mpx
lcp1
pbx1a
pbx3b
pu.1
lyz
Developmental stage (hpf)
e
flk1
fli1
tle1
tle2
Developmental stage (hpf)

FIG. 8

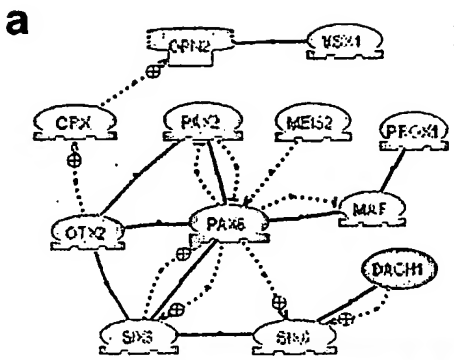


FIG. 11

Real-time RT- PCR and Profiling Analysis

Gene Name	Ratio (*clo*/WT)				
	Microarray			Real-Time PCR	
	15 hpf	30 hpf	48 hpf	30 hpf	48 hpf
meis2	1.12	1.25	0.48	1.76	0.15
prox1	0.98	1.27	0.29	1.92	0.33
vsx1	0.81	0.70	0.16	1.42	0.47
dachc	1.01	0.68	0.58	1.09	0.60
opn1sw2	0.83	0.77	0.17	1.06	0.46

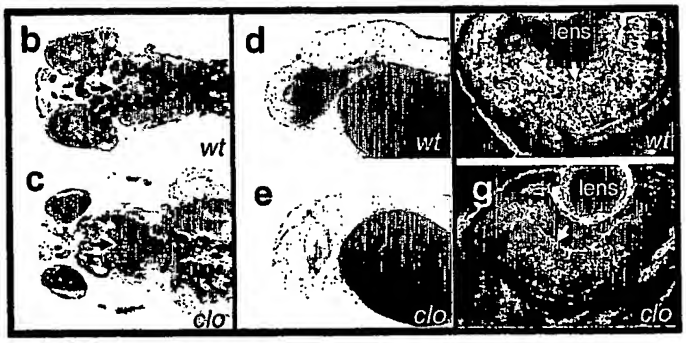


FIG. 12